DOCUMENT RESUME

ED 408 340 TM 026 593

AUTHOR	Wang, Wen-chung
TITLE	Estimating Rater Severity with Multilevel and Multidimensional Item Response Modeling.
SPONS AGENCY	Taiwan National Science Council, Taipei.
PUB DATE	[97]
NOTE	30p.
CONTRACT	85-2511-S-002-003
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	College Entrance Examinations; *Constructed Response; Foreign Countries; Interrater Reliability; *Item Response Theory; *Mathematical Models; *Test Items
IDENTIFIERS	Rater Effects; Taiwan

ABSTRACT

Traditional approaches to the investigation of the objectivity of ratings for constructed-response items are based on classical test theory, which is item-dependent and sample-dependent. Item response theory overcomes this drawback by decomposing item difficulties into genuine difficulties and rater severity. In so doing, objectivity of ability estimates is achieved, even though objectivity of ratings is poor. However, most item response models are too rigid to fit complexity of rater severities. Also, other types of items in the same test are excluded when estimating rater severities. These problems are addressed in this study. Several advanced models are proposed to explore severity changes over items and within items. In addition, multilevel and multidimensional models are formed to incorporate both multiple-choice items and constructed-response items in the test to increase estimating accuracy and model fit. The proposed models are made possible by a newly developed item response model, the multidimensional and multilevel random coefficient multinomial logit model. A real data set from the biology subject of the 1995 Joint College Entrance Examination in Taiwan was analyzed to demonstrate the advantages of this approach. (Contains 5 tables, 6 figures, and 35 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 340

Estimating Rater Severity with Multilevel and Multidimensional Item Response Modeling

Wen-chung Wang

National Chung-Cheng University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Correspondence:

Department of Psychology

National Chung-Cheng University

Chia-Yi, Taiwan

Phone: 886-5-2720411 ext. 6430

Fax: 886-5-2720857

E-mail: psywcw@ccunix.ccu.edu.tw

Acknowledgments:

This study was supported by resources of Contract Number 85-2511-S-002-003 from the National Science Council, Taiwan.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Wen-chung Wang

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

TM026593

Abstract

Traditional approaches to investigation of objectivity of ratings for constructed-response items are based on classical test theory, which is item-dependent and sample-dependent. Item response theory overcomes this drawback by decomposing item difficulties into genuine difficulties and rater severities. In so doing, objectivity of ability estimates is achieved, even though objectivity of ratings is poor. However, most item response models are too rigid to fit complexity of rater severities. Also, other types of items in the same test are excluded when estimating rater severities. These problems are resolved in this study. Several advanced models are proposed to explore severities changes over items and within items. In addition, multilevel and multidimensional models are formed to incorporate both multiple-choice items and constructed-response items in the test to increase estimating accuracy and model fit. The proposed models are made possible by a newly developed item response model, the multidimensional and multilevel random coefficients multinomial logit model. A real data set from the biology subject of the 1995 Joint College Entrance Examination in Taiwan was analyzed to demonstrate the advantages of this approach.

Key words: Rasch model, rater severity, objectivity of ratings, objectivity of ability estimates, multilevel item response models, multidimensional item response models.

Multiple-choice (MC) items have dominated large-scale testing for many years. The administrative convenience, economic advantages, and objective rating make MC items automatic choice for many test. In fact, objectivity of ratings is not inherited. For instance, when the Committee of the Joint College Entrance Examination in Taiwan announced the so called correct answers for the MC items, some of them were challenged and finally modified. Fortunately, through discussions and consensus, objectivity of rating for the MC items is reasonably achieved.

On the other hand, the Committee never announces scoring rubrics for the constructed-response (CR) items in the examination, not to mention the correct or reference answers. Usually, each CR item is judged by two raters independently. Some rating variations are usually found. For such an important examination, examinees may be accepted or rejected by colleges simply due to a difference of one point. Consequently, the impacts of rating variations on the examinees are very substantial.

Traditionally, consistency of ratings is treated as an index of objectivity of ratings. If the consistency is low, the objectivity is poor, and vice versa. The Pearson correlation is widely used to assess interrater reliability. Percentages of agreement are also used to account for consistency of ratings. In recent years, generalizability theory has been applied to assess the generalizability of scores across raters, items, or conditions (Abedi & Baker, 1995; Lavingueur, Tremblay, & Saucier, 1993; Longford, 1994; Marcoulides, 1994; McWilliam & Ware, 1994).

Although these methods seem reasonable, there are some drawbacks. First, even if the two ratings are identical or the correlation is perfect, it does not necessarily mean that the examinees are given what they deserve. These two raters might consistently give a faulty score, either too high or too low. Therefore, the two ratings are consistently incorrect. Objectivity of ratings is thus poor. Second, both the correlation or generalizability theory are based on classical test theory. The scales in classical test theory are assumed to be interval. However, raw scores or their linear transformations are not interval per se (Wright & Stone, 1979).

Objectivity of ratings refers to as the degree of agreement between given scores and deserved scores. So called deserved scores are the scores should be given in

theory. If the agreement is perfect, the objectivity is perfectly achieved. If the given scores are lower than the deserved scores, the examinees are severely treated. On the other hand, if the given scores are higher than the deserved scores, the examinees are leniently treated. The problem here is how to acquire the deserved scores. In practice, the scores given by rating experts can be treated as the deserved scores, because they are believed to have thorough understandings of the construct assessed and the examinees' responses. The given scores are derived from general raters. If the given scores are close to the deserved scores, objectivity of ratings is warranted.

In practice, some artificial responses or real ones sampled from the data set are judged by experts in advance. General raters are then asked to judge these responses. If their ratings are consistent with those of the experts, the raters are viewed as successfully "anchored" on the experts and well qualified. Otherwise, they need further training. This anchoring process can take place whenever needed, such as once a day. Unfortunately, because the experts can only judge a small portion of the responses, evaluation of objectivity of ratings is limited.

Even though objectivity of ratings is achieved through whatever quality control system, another impotent issue arises afterwards: How can we objectively estimate examinees' ability levels? As we know, the ratings of MC items are generally considered objective, but the ability estimates based on classical test theory, usually raw scores or their linear transformations, are item-dependent, thus, not objective. Item response theory resolves this problem by incorporating an item characteristic curve (ICC) between ability levels and probabilities of item responses. If data fit the item response models, the ability estimates and the item difficulty estimates are mutually independent. Consequently, the ability estimates are objective.

For items that are not objectively scored, such as CR items, how can objectivity of ability estimates be achieved? The key point is to discompose item difficulty into two subfacets: genuine difficulty and rater severity. For MC items, the item difficulty is equivalent to the genuine difficulty because no raters are involved. However, for CR items where raters are involved, the item difficulty contains not only the genuine difficulty but also the difficulty caused by specific raters, which is referred to as rater severity. By definition, if rater n's severity is positive, the items she

judged become more difficult. If her severity is negative, the items become easier. If her severity is zero, the items stay unchanged. Likewise, if data fit the models, the ability estimates are independent of both the genuine difficulties and the rater severities, thus are objectively estimated.

Research works using item response theory to investigate rater severity have been accumulated, such as Engelhard (1992, 1994, 1996), Lunz & Stahl (1990a, b), Lunz, Stahl, Wright, & Linacre (1989), Lunz, Wright, & Linacre (1990), Lunz, Wright, Stahl, & Linacre (1989), Wang & Wilson (1996), Wilson & Wang (1995), to name a few. There are two major shortcomings for most of these research works. One is oversimplification on modeling of rater severity. Only one parameter was usually assigned to each rater to represent the severity. Rater severity was not allowed to interact with items. In practice, raters may show different severities across items or even within items. The other is oversimplification on estimation procedures. Only CR items were used to estimate the ability, the genuine difficulty, and the rater severity parameters. Items in the same test, such as MC items, were put aside. However, being in the same test, these MC items possess some information about the parameters. Therefore, the information should be taken into account.

In this study, some advanced models for rater severity are proposed. Rater severities are allowed to be different across items and within items. Besides, multidimensional and multilevel models are proposed to incorporate information of the MC items in the same test to improve model fit. The advanced modeling is made possible through a newly developed item response model, the multidimensional and multilevel random coefficients multinomial logit model (M^2RCML), which I briefly introduce in the latter section. Finally, a small real data set from the biology subject of the 1995 Joint College Entrance Examination in Taiwan was analyzed to demonstrate the advantages of the advanced modeling.

Some Advanced Models for rater Severity

In most of traditional item response models, only two facets are involved, examinee's ability and item difficulty. In order to achieve "specific objectivity", these two facets should not interact with each other (Rasch, 1960/1980; Wright & Masters, 1982). The item difficulty facet can actually be partitioned into several facets when needed. For example, in CR items where raters are involved, the item

difficulty facet can be partitioned into genuine item difficulty and rater severity. These two subfacets as well as the ability facet are in the same unit, the logit.

A three-way factorial design can be used to depict the relations among these facets. Let the examinee's ability, the genuine difficulty, and the rater severity be treated as three independent variables, and the item responses as a dependent variable. As usual, the ability facet and the genuine difficulty facet should not interact with each other, and the ability facet and the rater severity facet should not interact with each other, either, because the two subfacets are partitioned from the item difficulty facet. However, the genuine difficulty facet is allowed to interact with the rater severity facet.

Most of the earlier research constrained these two subfacets to be independent, hence, reduced flexibility of modeling of rater severity. In the following, this constraint is released. The liberation has two advantages. The data analysts can gain deep understanding about patterns of severity changes across and within items. Quality control systems can accordingly be implemented to increase objectivity of ratings. In addition, more accurate modeling can improve model fit, including objectivity of ability estimates. Suppose raters did exhibit different severities across or within items, using traditional approaches cannot fully capture the complexity. The data would not fit the simpler model well, which in turn reduces the estimation accuracy.

Dichotomous Items

Let θ_n denote examinee n's ability, and δ_i denote item i's difficulty. Let p_{ni0} denote the probability of an incorrect answer (scored as 0) of examinee n to item i, and p_{ni1} denote that of a correct answer (scored as 1) of examinee n to item i. In the Rasch model,

$$log\ (p_{ni1} / p_{ni0}) = \theta_n - \delta_i, \qquad (1)$$

where $log\ (p_{ni1} / p_{ni0})$ is a log-odd with a unit of logit. In the above equation, only two facets are involved, the ability facet and the item difficulty facet. In addition, θ_n does not depend on item (no subscript i), and δ_i does not depend on examinee (no subscript n). That is, θ_n and δ_i are mutually independent.

Suppose these items are judged by raters and they may express different severities, the item difficulty can then be partitioned into the genuine difficulty and the rater severity. Let there be I items, indexed $i = 1, \ldots, I$, and R raters, indexed $r = 1, \ldots, R$. We can rearrange items by raters to form rater-items, indexed $k = 1, \ldots, K$. For example, if there are 10 items and two raters, with each item judged by these two raters independently. In such a case, 20 rater-items are formed. Each rater-item has one difficulty parameter, resulting in 20 difficulty parameters altogether can be estimated at most. Then, δ_i in Equation (1) becomes ξ_k, where ξ denotes parameters of the rater-items. We can use the Rasch model to estimate all the 20 rater-item difficulty parameters, by simply rearranging the 10 items with each of two ratings into 20 rater-items physically. However, in so doing, the genuine difficulties and the rater severities are confounded.

This problem can be easily resolved by constraining:

$$\xi_k = \delta_i + \rho_r, \tag{2}$$

where δ_i is item i's genuine difficulty, and ρ_r is rater r's severity. Note that in Equation (2) δ_i and ρ_r are independent (no subscript r for items and no subscript i for raters). With this modeling, Equation (1) becomes

$$log\ (p_{ni1} / p_{ni0}) = \theta_n - (\delta_i + \rho_r). \tag{3}$$

The genuine difficulties and the rater severities are separated. If each item is judged once by a distinct rater, the two subfacets are totally confounded and no rater severity can be identified. For dichotomous items, subject to model identification, we cannot estimate both the genuine difficulties, δ_i, and the rater severities across items, ρ_{ir}. However, for polytomous items, rater severities across or within items can be successfully estimated. I shall come to this issue in the following section.

Polytomous Items

For polytomous items, Equation (1) can be extended into

$$\log\ (p_{nij} / p_{nij-1}) = \theta_n - \delta_{ij}, \tag{4}$$

where p_{nij} and p_{nij-1} denote the probabilities of scoring j and j-1, respectively, of examinee n to item I; θ_n denotes examinee n's ability; δ_{ij} denotes step j's difficulty of

item i. If an item is scored 0, 1, ... , J, a set of J step difficulties can be estimated. This is the partial credit model (Masters, 1982).

If raters are involved in polytomous items, similar to dichotomous items, rater-items can be formed and indexed $k = 1, \ldots, K$. Although we can estimate a set of J step difficulties for each rater-item, the genuine difficulties and the rater severities are confounded. As in dichotomous items, this problem can also be resolved by constraining

$$\xi_{kj} = \delta_{ij} + \rho_{rj}, \tag{5}$$

where δ_{ij} is item i's genuine difficulty, and ρ_{rj} is rater r's severity, at the j step. Note that δ_{ij} and ρ_{rj} are independent, thus the genuine item difficulties and the rater severities are separated.

In the partial credit model, it can be shown that through reparameterization δ_{ij} can be decomposed into an overall difficulty and several threshold difficulties:

$$\delta_{ij} = \delta_{i\cdot} + \tau_{ij}, \tag{6}$$

where $\delta_{i\cdot}$ is the center of δ_{ij} and referred to as the overall difficulty of item i; τ_{ij} is the deviance of δ_{ij} to $\delta_{i\cdot}$ and referred to as the threshold difficulty of step j of item i. Similarly, the rater severities can be reparameterized into an overall severity and several threshold severities:

$$\rho_{rj} = \rho_{r\cdot} + \eta_{rj}, \tag{7}$$

where $\rho_{r\cdot}$ is the overall severity of rater r, and η_{rj} is the threshold severity of step j of rater r. With this parameterization, Equation (4) becomes

$$\log (p_{nij} / p_{nij\text{-}1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r\cdot} + \eta_{rj}). \tag{8}$$

In the rating scale model (Andrich, 1978), the threshold difficulties are constrained to be identical across items:

$$\delta_{ij} = \delta_{i\cdot} + \tau_j, \tag{9}$$

where τ_j does not depend on items. Likewise, the threshold severities can be constrained to be independent of raters:

$$\rho_{rj} = \rho_{r\cdot} + \eta_j. \tag{10}$$

Therefore, a reduced model is formed:

$$\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_r + \eta_j). \quad (11)$$

In this model, raters are supposed to express different overall severities but identical threshold severities. If the model fits the data well, it will be preferred to the more complicated model, Equation (8).

The above model, Equation (11), can be further reduced by constraining all the threshold severities to be zero and leads to

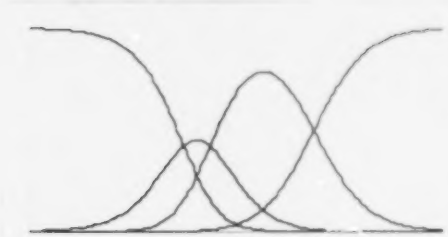
$$\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_r). \quad (12)$$

It is the model that was widely used in the literature. Raters are expected to express different overall severities but no threshold severities.

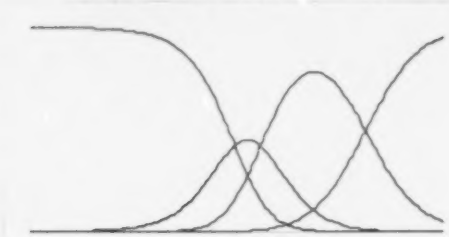
Figure 1 shows how the overall severities and the threshold severities affect the ICCs. In Figure 1a, both the overall and the threshold severities are zero, treated as the reference. In Figure 1b, the overall severity is positive but all the threshold severities are zero. Comparing Figures 1a and 1b, we find that the patterns of the ICCs are identical except the location of Figure 1b shifting to the right, which means that the item becomes more difficult. Conversely, when the overall severity is negative, the location of the ICC shifts to the left, as shown in Figure 1c, meaning that the item becomes easier. Adding an overall severity changes the overall difficulty, whereas the threshold difficulties are not affected. Consequently, only the location of the ICCs shifts and the pattern remains unchanged.

From Figures 1d to 1f, the overall severities are zero, but the threshold severities do exist. In Figures 1d and 1e, the first threshold severities are positive and negative, respectively, with both the second threshold severities being zero. On the contrary, in Figures 1f and 1g, the second threshold severities are positive and negative, respectively, with both the first threshold severities being zero. Treating Figure 1a as reference, we find that the overall difficulties (center of the ICCs) of Figures 1d to 1g remain unchanged. However, the patterns differ. In sum, the overall severities change the locations, and the threshold severities change the patterns.

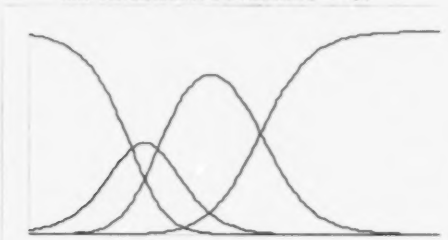
1a. No rater severity



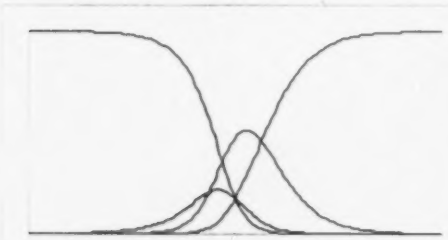
1b. Overall severity > 0,
all threshold severities = 0.



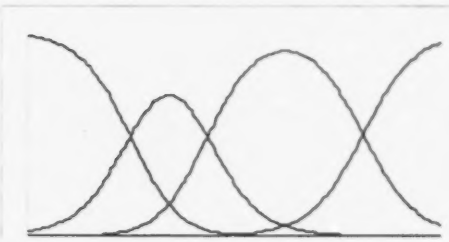
1c. Overall severity > 0,
all threshold severities = 0.



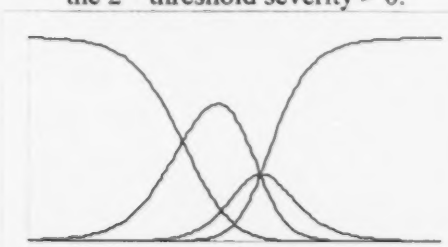
1d. Overall severity = 0,
the 1st threshold severity > 0.



1e. Overall severity = 0,
the 1st threshold severity < 0.



1f. Overall severity = 0,
the 2nd threshold severity > 0.



1g. Overall severity = 0,
the 2nd threshold severity < 0.

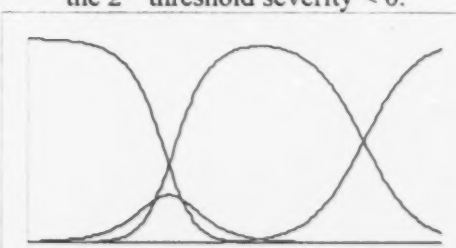


Figure 1. Influences of rater severity on item characteristic curves

In all the above models, raters severities are constrained to be independent of items. In some cases, they are allowed to interact. We simply add a subscript i to the overall severities and the threshold severities, which changes Equation (8) to:

$$\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r+i} + \eta_{rij}). \quad (13)$$

Unfortunately, this model is not identifiable because the number of parameters is greater than the number of informations. We can constrain the threshold severities to be independent of items, resulting in the following identifiable model:

$$\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r+i} + \eta_{rj}). \quad (14)$$

In this model, raters are expected to express different overall severities across items, but the threshold severities are constant across items.

Equation (14) can be reduced to a simpler model by constraining the threshold severities to be constant across raters, resulting in:

$$\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r+i} + \eta_{j}). \quad (15)$$

In this model, although raters express different overall severities across items but their threshold severities are constant across both raters and items.

We further simplify the above model by constraining all the threshold severities to be zero, which leads to:

$$\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r+i}). \quad (16)$$

Now, raters are expected to express different overall severities across items, but their threshold severities are all zero. The six identifiable models are summarized in Table 1. Note that models 1, 2, and 3 are nested. Models 4, 5, and 6, models 1 and 4, models 2 and 5, models 3 and 6, are nested, too.

Table 1. Some item response models for rater severities

Model	Equation	Model Hierarchy
1	8: $\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_{r\cdot} + \eta_{rj})$	Submodel of model 4
2	11: $\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_{r\cdot} + \eta_j)$	Submodel of models 1 and 5
3	12: $\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_{r\cdot})$	Submodel of models 2 and 6
4	14: $\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_{ri} + \eta_{rj})$	
5	15: $\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_{ri} + \eta_j)$	Submodel of model 4
6	16: $\log (p_{nij} / p_{nij-1}) = \theta_n - (\delta_i + \tau_{ij} + \rho_{ri})$	Submodel of model 5

Multidimensional and Multilevel Modeling

It is very common that a test contains several item formats, such as MC items and CR items. When estimating rater severities, usually only CR items are used. The MC items in the same test are put aside. In fact, the MC items provide some information about examinees' ability, which in turn can be used to improve estimation accuracy of rater severities and other parameters. When there are only a few CR items, or when raters judge only a few CR items, estimation of rater severities based on the CR items only would be very imprecise. In such a case, information of the MC items can really help.

Applying traditional item response models to incorporate information of both item formats, we have to treat both item formats as unidimensional. In so doing, we run into another argument: Are these two item formats really unidimensional? There is no easy answer. Usually, test developers adopt different item formats to tap different dimensions. For instance, MC items usually aim at low level abilities, such as knowledge and understanding, whereas CR items aim at high level abilities, such as application and problem-solving. If so, treating both item formats as unidimensional will contaminate the underlying dimensions. Even worse, the individual dimensions disappear. Users of traditional item response models are forced to either discard the MC items so that the underlying dimensions are not confounded but estimation accuracy will be imprecise, or treat both item formats as unidimensional to improve estimation accuracy but contaminate the dimensions.

There are two alternatives which can resolve this problem: multidimensional modeling and multilevel modeling. If we have reasons to believe that different item

formats tap different dimensions, we should not treat them as unidimensional but multidimensional. Thus, multidimensional item response models are needed, where individual dimensions are simultaneously estimated. This approach has three advantages. First, the above dilemma of using unidimensional models disappears. Second, the individual dimensions are reserved and each examinee has two ability estimates, one for each dimension. Third, because one dimension can provide some collateral information on the other, estimation accuracy can be improved, especially when the two dimensions are highly correlated. All we need is a multidimensional item response model.

Multilevel modeling results from recognition of multilevel structures of data. In the literature, multilevel models have several names. For examples, in the social sciences, they are referred to as hierarchical linear models (Bryk & Raudenbush, 1992; Lindley & Smith, 1972) or multilevel linear models (Goldstein, 1977); in biometrics, as mixed models or random effect models (Laird and Ware, 1982); in econometrics, as random coefficients regression models (Rosenberg, 1973); in statistics, as covariance component models (Dempster, Rubin, & Tsutakawa, 1981).

Most multilevel models are linear and based on classical test theory. It is desirable to develop a multilevel model which is based on item response theory. Such a model not only has the advantages of multilevel over unilevel but also the advantages of item response theory over classical test theory. At the first level of the multilevel item response model, a conditional item response model is formed. At the second level, the ability parameters are regressed on some predictors, such as examinees' background variables. The parameters at the two levels are jointly estimated. The same idea can be applied to investigate rater severities. The raw scores or ability estimates of MC items can be treated as the predictor at the second level.

Both multidimensional modeling and multilevel modeling can incorporate information of MC items into estimation of abilities and rater severities. In multidimensional modeling, the original responses of MC items are used, whereas in multilevel modeling, the raw scores or the ability estimates of MC items are used. In the former, these two dimensions based on MC and CR items, respectively, are

simultaneously estimated. In the latter, the abilities based the MC items are estimated first, then put into the second level to help estimate the abilities and the rater severities based on the CR items. Therefore, in theory, the estimation accuracy of multidimensional modeling should be greater than that of multilevel modeling.

To utilize these two alternatives as well as the advances models for rater severities, a multidimensional and multilevel item response model is needed. A newly developed item response model can meet the demand. It is briefly introduced in the following section.

The Multidimensional and Multilevel Random Coefficients Multinomial Logit Model (M^2RCML)

The M^2RCML (Adams, Wilson, & Wang, in press; Adams, Wilson, & Wu, in press; Wang, 1994) is a multidimensional and multilevel extension of the random coefficients multinomial logit model (Adams and Wilson, 1996). The M^2RCML has two levels: a between student level and a within student level. At the between student level, a population model $f_\theta(\theta;\alpha)$ is formed, where θ is a vector of latent variables and α is a set of parameters that characterize the distribution of θ. The population model describes the between student variation in the latent variables. At the within student level, a conditional item response model $f_x(\mathbf{x};\xi \mid \theta)$ is formed, where $\mathbf{x}$ is a vector of observation on items, ξ is a vector of parameters that describe those items, and θ is a vector of latent variables. The conditional item response model describes the probability of observing a set of item responses conditioned on the level of an individual on the set of latent variables.

The Population Model

Regarding the D-dimensional distribution of $\theta = (\theta_1,...,\theta_D)'$, a multivariate normal distribution is assumed to simplify estimation procedure. In addition to the θ, we have a vector of $\mathbf{u}$ observed background characteristics. For examinee n, the background characteristics are $y_{n1}, ..., y_{nu}$, which are collected in to a vector $\mathbf{Y}_n = (1, y_{n1}, ..., y_{nu})'$. Treating the $\mathbf{u}$ observed background characteristics as predictors of the θ latent variables, we have $\theta_n = \Gamma\mathbf{Y}_n + \mathbf{E}_n$, where Γ is a $d \times u$ matrix of regression

coefficients, $\mathbf{E}_n \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$, and $\mathbf{\Gamma Y}_n$ and $\mathbf{E}_n$ are independent, as assumed in the usual regression models.

The Conditional Item Response Model

Suppose a set of D latent traits underlie the examinees' test performances and the examinees' positions are denoted $\theta = (\theta_1, ..., \theta_D)'$. Let there be I items indexed i = 1, ... , I, and K_i response categories in item i indexed k = 1, ... , K_i. A response in category k of item i is scored b_{ikd} on dimension d (the scoring rubrics are known a priori). The scores across D dimensions can be collected into a column vector $\mathbf{b}_{ik}$ = $(b_{ik1}, ..., b_{ikD})'$, then into a scoring sub-matrix for item i, $\mathbf{B}_i = (\mathbf{b}_{i1}, ..., \mathbf{b}_{iK_i})'$, and then into a scoring matrix $\mathbf{B} = (\mathbf{B}_1', ..., \mathbf{B}_I')'$ for the whole test.

Let $\xi = (\xi_1, ..., \xi_p)'$ denote a vector of p free item parameters. Let a design vector $\mathbf{a}_{ik}'$ denote a linear combinations of ξ corresponding to response category k of item i. They are denoted by a design matrix $\mathbf{A} = (\mathbf{a}_{11}', \mathbf{a}_{12}', ..., \mathbf{a}_{1k_1}', \mathbf{a}_{21}', ..., \mathbf{a}_{2k_2}', ..., \mathbf{a}_{Ik_I}')'$ for the whole test. Let an indicator variable X_{nik} denote as

$$X_{nik} = \begin{cases} 1 & \text{if response of examinee } n \text{ to item } i \text{ is in category } k, \\ 0 & \text{otherwise.} \end{cases}$$

Under the M^{2}RCML model, the probability of a response in category k of item i for examinee n is expressed as

$$f(X_{nik} = 1, \mathbf{A}, \mathbf{B}, \xi \mid \theta_n) = \frac{\exp(\mathbf{b}_{ik}' \theta_n + \mathbf{a}_{ik}' \xi)}{\sum_{u=1}^{K_i} \exp(\mathbf{b}_{iu}' \theta_n + \mathbf{a}_{iu}' \xi)}.$$

A marginal maximum likelihood estimation with EM algorithm (Bock & Aitkin, 1981) for the model is developed and implications and applications are also shown. Interested readers are referred to Adams, Wilson, & Wang (in press), Adams, Wilson, & Wu (in press), Wang (1994), and Wang, Wilson, & Adams (1995, in press) for details.

Real Data Analyses

A real data set from the biology subject of the 1995 Joint College Entrance

Examination in Taiwan was analyzed to demonstrate various item response models for rater severities. The test booklet is consisted of 20 MC items, 20 multiple MC items, and 5 CR items, with four sub-items in each CR item. Each sub-item was scored 0, 1, or 2, resulting in a total score from 0 to 8 for each CR item. Each examinee was judged by two raters independently. If a difference of more than two points was found for each CR item, a third rating will be given.

For simplicity, only the 20 MC items and the 5 CR items, 414 examinees, and 15 raters were used. On the average, each rater judged about 50 examinees on each CR item. Although the data set is small, it does not limit generalization and implication of the modeling.

Basic Analyses

Since each CR item was judged by two raters, an average of these two scores is given to the examinees. Figure 2 presents the means and the standard deviations of the CR items for these 414 examinees. The mean of the third item is the highest, 6.53, and that of the forth item is the lowest, 2.34. Regarding the standard deviations, the third item is the lowest, 1.65, and the others are between 2.08 and 3.05. Therefore, it seems that the third item is the easiest, the forth item is the most difficult, and the ratings of the third items is the least dispersed.

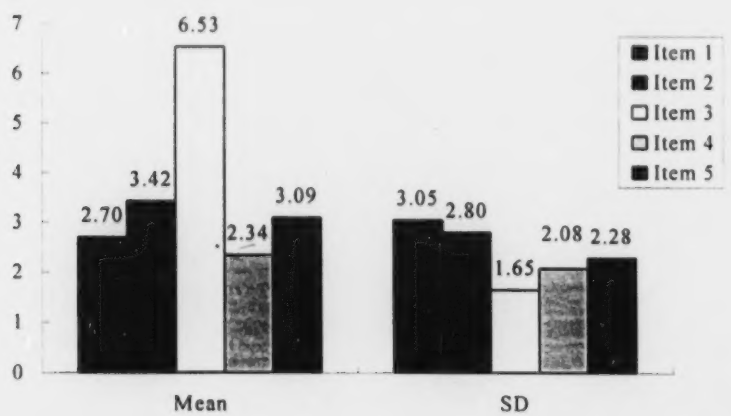


Figure 2. The means and the standard deviations of the CR items

Percentages of agreement can be used to represent interrater consistency. Since the maximum possible score of each CR item is 8, the differences of the two ratings can be at most -8 or 8. Table 2 shows the differences and the percentages of the two

ratings for each CR item. For percentages of perfect agreement, the third item and the first item are the highest, 94.52% and 93.00%, respectively. The percentages of more than a ±2 point difference are less than 1% for every item. Generally speaking, the distributions of the differences of the two ratings for the CR items are symmetric. No systematic biases are found in the two ratings, say, the first ratings are systematically lower or higher than the second ratings. The correlations of the two ratings for the items are between .93 and .98. All these results point out satisfactory interrater consistency.

Table 2. The frequencies of the differences and the percentages of the two ratings for the 5 CR items

Item Difference	1	2	3	4	5
-4	2 (.48)*	1 (.24)	0 (.00)	0 (.00)	1 (.24)
-3	0 (.00)	0 (.00)	0 (.00)	0 (.00)	0 (.00)
-2	10 (2.42)	9 (2.17)	18 (4.35)	8 (1.93)	21 (5.07)
-1	3 (.72)	8 (1.93)	2 (.48)	26 (6.28)	33 (7.97)
0	385 (93.00)	371 (89.61)	390 (94.52)	332 (80.19)	308 (74.40)
1	2 (.48)	7 (1.69)	0 (.00)	39 (9.42)	24 (5.80)
2	12 (2.90)	18 (4.35)	5 (1.21)	6 (1.45)	24 (5.80)
3	0 (.00)	0 (.00)	0 (.00)	3 (.72)	1 (.24)
4	0 (.00)	0 (.00)	0 (.00)	0 (.00)	2 (.48)

Values in parentheses are percentages.

Figure 3 shows the mean ratings of the CR items given by the raters. On the average, Rater 4 gave the highest scores, with a mean of 4.15; Rater 11 gave the lowest scores, with a mean of 3.09. An instant thought might be that Rater 4 is the most lenient and Rater 11 is the most severe. A score given by Rater 11 is 1.06 points less than that by Rater 4, on the average. For 5 items, a difference of 5.30 points could be found. This difference, of course, could affect the examinees significantly in such a competing entrance examination.

Rater 4 and Rater 11 would have felt very uncomfortable about this conclusion. Rater 11 might argue that the performances she judged are worse than those in general and thus the ratings she gave were actually in consistent with the scores should be. Therefore, it should not be problematic that her ratings are lower than others. There is no easy way to approve or disapprove her argument, because examinees' abilities

and rater severities are confounded.

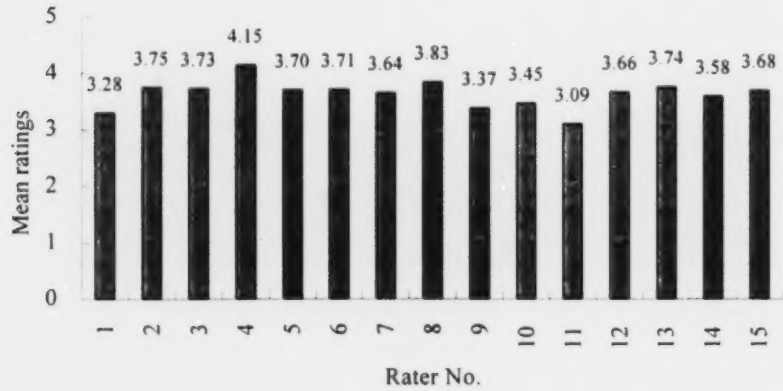


Figure 3. The mean ratings of the 15 raters on the 5 CR items

Item Response Modeling

In the following, I first apply traditional approach, unilevel and unidimensional item response modeling, then move to multilevel item response modeling and multidimensional item response modeling. The analyses were made possible by the software MATS (Wu, Adams, & Wilson, 1995).

Unilevel and Unidimensional Modeling

Based on the 5 CR items and 15 raters, with the 20 MC items excluded, the loglikelihood deviances of these six models are shown in Table 3. For Model 1, each rater has one overall severity and seven threshold severities, resulting in 14 overall severity parameters (one treated as reference for model identification) and 105 (= 7 × 15) threshold severity parameters. However, some raters did not give some particular scores and thus four parameters are not estimated, which leads to 115 rater severity parameters altogether. The loglikelihood deviance of Model 1 is 11041.63.

Model 2 is a submodel of Model 1, by constraining all the threshold severity parameters to be identical across raters, which leaves out additional 94 parameters. Model 3 is a submodel of model 2, by constraining all the threshold severity parameters to be zero, which leaves out additional 7 parameters. To compare these three models, the usual loglikelihood ratio test can be applied. Similarly, Models 4, 5, and 6 are nested. As stated above, Model 1 is a submodel of Model 4; Model 2 is a submodel of Model 5; Model 3 is also a submodel of Model 6. Figure 4 shows the loglikelihood ratio tests of these six models. For Models 1 and 2, the difference of

BEST COPY AVAILABLE

the loglikelihood deviance is $\Delta G^2 = 89.65$, with degrees of freedom of $\Delta RP = 94$ and a p-value of .61, which indicates that these two models are not statistically significant. For parsimony, Model 2 is preferred. The other loglikelihood ratio tests can be done in the same way.

Table 3. The loglikelihood deviances and numbers of parameters of Models 1 to 6

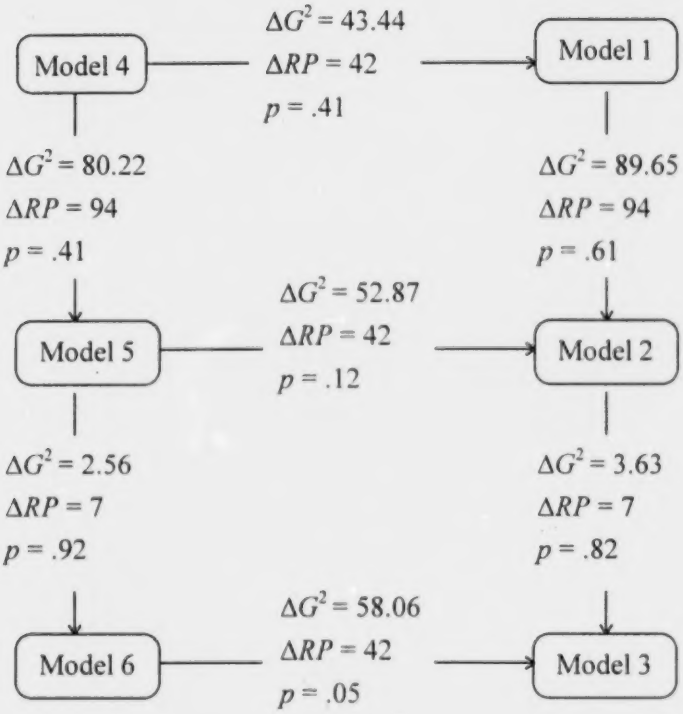
Model	Equation	Loglikelihood Deviance: G^2	# Rater Para.: RP
1	8: $\log (p_{nij} / p_{nij\text{-}1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r\cdot} + \eta_{rj})$	11041.63	115
2	11: $\log (p_{nij} / p_{nij\text{-}1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r\cdot} + \eta_{j})$	11131.28	21
3	12: $\log (p_{nij} / p_{nij\text{-}1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r\cdot})$	11134.91	14
4	14: $\log (p_{nij} / p_{nij\text{-}1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r\cdot i} + \eta_{rj})$	10998.19	157
5	15: $\log (p_{nij} / p_{nij\text{-}1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r\cdot i} + \eta_{j})$	11078.41	63
6	16: $\log (p_{nij} / p_{nij\text{-}1}) = \theta_n - (\delta_{i\cdot} + \tau_{ij} + \rho_{r\cdot i})$	11076.85	56

For models that are not nested, such as Models 1 and 5, Akaike's (Akaike, 1977) information criterion, *AIC*, can be applied, which is defined as:

$$AIC = G^2 + 2TP,$$

where G^2 is the loglikelihood deviance, and TP is the total number of parameters estimated. In this particular data set, TP is equal to RP plus 41, because additional 39 item parameters and 2 person parameters (mean and variance) were estimated.

From both the loglikelihood ratio tests and the *AIC*s, we find that Model 3 is the most parsimonious model. Therefore, only a single overall severity is needed for each rater.



Note: The arrows indicate nested models.

Figure 4. The loglikelihood ratio tests of Model 1 to Model 6

Multilevel Modeling

In the previous unilevel item response modeling, only the CR items were used and the MC item were discarded. In fact, the MC items and the CR items are moderately correlated as shown in Table 4. Therefore, information from the MC items could help improve estimation accuracy of examinees' abilities, item difficulties, and rater severities.

Table 4. The correlations of the raw scores between the MC items and the CR items

	Total Scores of the MC Items	CR Item 1	CR Item 2	CR Item 3	CR Item 4	CR Item 5
CR Item 1	0.77					
CR Item 2	0.77	0.76				
CR Item 3	0.25	0.23	0.21			
CR Item 4	0.59	0.55	0.55	0.20		
CR Item 5	0.62	0.61	0.59	0.23	0.53	
Total Scores of the CR Items	0.83	0.88	0.87	0.43	0.75	0.79

In multilevel modeling, either the total raw scores or the ability estimates were treated as the second-level predictor. In so doing, an additional parameter, a regression coefficient for the predictor, was estimated. As stated in the previous section, Model 3 is preferred, therefore, only Model 3 was adopted in the first-level modeling. At the second-level, in one hand, the raw scores of the MC items were treated as the predictor, which is referred to as Model 7. On the other hand, the ability estimates of the MC items derived from the Rasch model were treated as the predictor, which is referred to as Model 8.

As shown in Table 5, these two models have the loglikelihood deviances of 10976.62 and 11056.45, respectively, and *AIC*s of 11088.62 and 11168.45, respectively. These two *AIC*s are 156.29 and 76.46 less than that of Model 3. Therefore, these two models are significantly better than Model 3. This demonstrates that the multilevel modeling indeed helps increase model fit. Note also that Model 7 has a better fit than Model 8. However, Model 7 is based on the raw scores, which are not in interval scale. The scale of the second-level predictors in multilevel modeling is assumed to be interval. Hence, Model 7 violates this assumption. In contrast, the ability estimates derived from the Rasch model are interval. The assumption is sustained in Model 8.

Multidimensional Modeling

In the previous multilevel modeling, the original item responses of the MC items, being aggregated to form either the total raw scores or the ability estimates, were invisible. Since the second-level predictor actually comes from items, it is reasonable to treat them as such. This calls for a multidimensional item response model, where the MC items are treated as one dimension and the CR items as another. The parameters of these two kinds of items are jointly estimated. As the correlation of the two dimensions gets larger, the benefit of the multidimensional modeling gets better.

For the MC items, the Rasch model was applied, whereas for the CR items, Model 3 was applied. This resulting multidimensional model is referred to as Model 9 hereafter. Since the MC items were treated as a dimension, there is no second-level predictor. Therefore, Model 9 is in fact a unilevel but multidimensional model. Of course, if other suitable predictors are found, multilevel and multidimensional models can be applied in a direct manner.

In Model 9, altogether 77 parameters were estimated with a loglikelihood deviance of 19915.36. This model cannot be directly compared to Models 3, 7, or 8, because it is based on both the CR and MC items, while the other models are based on the CR items only. For model comparison, a model for the MC items should be formed, which is referred to as Model 10. It has a loglikelihood deviance of 9646.18, with 21 parameters. The deviances and *AIC*s of Models 3 and 10 are summed. Similarly, those of Models 7 and 10, and Models 8 and 10 are summed. Model 9 can then be compared to the summed models. As shown in Table 5, Model 9 has the best model fit among them. According to Model 9, the correlation between the two dimensions is as high as .95, higher than that of raw scores between the MC and the CR items, .83. All of these results demonstrate that in general the multidimensional modeling is superior to both the unidimensional modeling and the multilevel modeling, especially when dimensions are highly correlated.

Table 5. Model comparisons of various models

Model	Description	Loglikelihood Deviance	# Parameters	*AIC*
3	Unilevel and unidimensional model: CR only	11134.91	55	11244.91
7	Multilevel and unidimensional model: Raw scores	10976.62	56	11088.62
8	Multilevel and unidimensional model: Ability estimates	11056.45	56	11168.45
9	Unilevel and multidimensional model	19915.36	77	20069.36
10	Unilevel and unidimensional model: MC only	9646.18	21	9688.18
3 + 10		20781.09	76	20968.54
7 + 10		20622.80	77	20776.8
8 + 10		20702.63	77	20856.63
11	Unilevel and unidimensional model: CR + MC	20022.76	74	20170.76

By applying the traditional item response models, the MC items and the CR items can be either calibrated separately or concurrently. What I mean concurrently here is that both types of items are put together, treated as unidimensional, and calibrated as such. This model, referred to as Model 11, has a loglikelihood deviance of 20022.76. It does not fit the data as well as Model 9. In addition, treating both types of items as unidimensional may run into a problematic argument: Are they really unidimensional?

Also, the individual dimensions are preserved in the multidimensional modeling whereas they are invisible and a only composite dimension is formed in the unidimensional modeling.

The overall severities derived from these models are very similar, especially when their standard errors are taken into account. Figure 5 displays the 95% confidence intervals of the overall severity parameters of these 15 raters in Model 9. Among them, those of Raters 6, 8, 10, and 13 contain zero. Note that Rater 15 does not have a standard error because the parameter is constrained for model identification. For those 14 raters, a chi-squared test was performed to check if the parameters are different from zero. It turned out that χ^2 equals to 209.23, with 14 degrees of freedom and a p-value less than .001. Consequently, not every raters expressed the same severities in this particular data set.

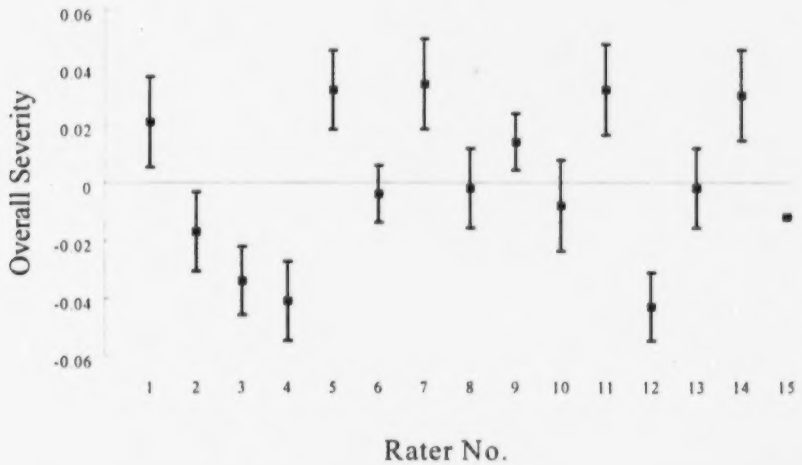


Figure 5. The 95% confidence intervals of the overall severities of the 15 raters

The overall severities of Raters 1, 5, 7, 9, 11, and 14 are positive, which means that items becomes more difficult if judged by these raters. In contrast, the overall severities of Raters 2, 3, 4, 12, and 15 are negative. Thus, items become easier when judged by them. Although the overall severities are significantly different from zero, they may not be practically important, because the severities are around the second decimal, between -.043 and .034, with a range of .077. The overall item difficulties and the standard errors for the 5 CR items are: .238 (.018), .029 (.020), -1.094 (.013), .604 (.024), and .224, respectively, with a range of 1.698. The range of the overall severities is only 4.5% of the range of the overall item difficulties. Generally

speaking, the raters had minute influences.

The finding is in concert with the basic analyses based on the raw scores, as shown in Table 2 where the two ratings are almost identical. The consistency may be due to the special structure of the CR items. As stated above, each CR item contains four subitems, with each judged in a three-point scale (0, 1, 2). In so doing, the scoring rubrics were well specified, thus, the rater consistency was foreseen.

As in Figure 3, we have shown that Rater 4 gave the highest scores and Rater 11 gave the lowest. The difference of these two mean ratings is 1.06 points. It seems that Rater 4 is the most lenient and Rater 11 is the most severe. However, this may not be the case because (a) they did not judge the same examinees, and (b) in such a small data set, it is unwise to assume that the two samples of examinees come from identical populations.

This problem can be resolved by estimating the rater severities. Figure 6 shows a scatter plot of the mean ratings and the overall severities of the 15 raters. The correlation is -.59, indicating that in general the higher the mean ratings are, the less the severities. However, we should be very cautious since the correlation is only moderate. This finding suggests that using the mean ratings to locate the severities is imprecise and questionable.

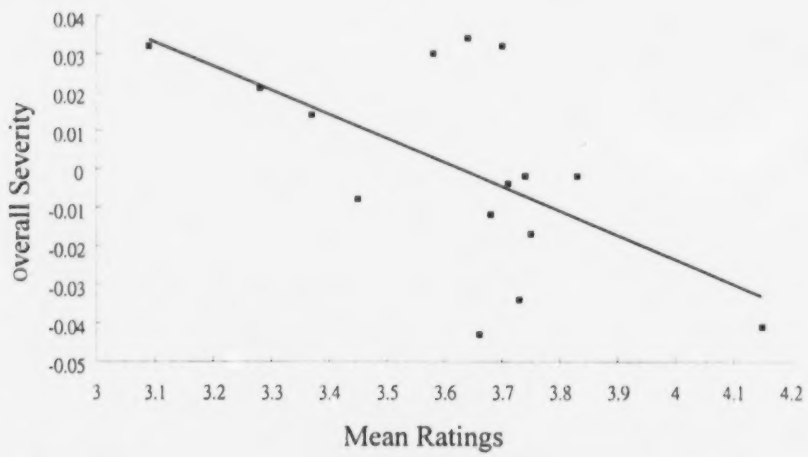


Figure 6. A scatter plot of the mean ratings and the overall severities of the 15 raters

Conclusions

Objectivity of ratings has long been the major problem of constructed-response items. Traditional approaches usually focus on consistency of ratings, such as percentages of agreement of ratings, the correlation for interrater reliability, or generalizability of scores across raters. However, consistency of ratings is not equivalent to objectivity of ratings, because two ratings may be consistently biased. In this study, objectivity of ratings is referred to as the degree of agreement between given scores and deserved scores. If the agreement is perfect, objectivity of ratings is perfectly achieved.

Even though objectivity of ratings is perfectly achieved, such as MC items, ability estimates based on classical test theory are item-dependent and thus not objective. Item response theory resolves this problem by incorporating ICCs between ability levels and probabilities of item responses. If data fit the item response models, the estimates of ability and item difficulty are mutually independent. Hence, objectivity of ability estimates is achieved.

For CR items where objectivity of ratings are usually not perfectly achieved, objectivity of ability estimates is still possible. Item difficulty can be decomposed into genuine difficulty and rater severity. Likewise, if data fit the models, the ability estimates are objective. However, most of earlier works using item response models have two major drawbacks: oversimplification on modeling of rater severity and oversimplification on estimation procedures.

In this study, these two problems are resolved. Several advanced models are proposed to investigate rater severities within items and across items. In so doing, complexity of rater severities is better monitored. In addition, multilevel and multidimensional modeling are applied to incorporate all information in a test to improve model fit as well as estimation accuracy of parameters. The various advanced models, the multilevel modeling, and the multidimensional modeling are made possible through the newly developed M^2RCML.

This study has four major implications. First, when a rating session just begins and only a few ratings are made, applying the multilevel and the multidimensional modeling can help increase estimates of rater severities. Once a large severity is found, remedy action can be implemented in time to maintain objectivity of ratings. Second, applying these advanced models can thoroughly detect variations of rater severities within items and across items as well as time, places, or situations. A better quality control system for ratings can then be set up. Third, use of these advanced models as well as the multilevel and the multidimensional modeling can better locate examinees' ability levels, even after the ratings are made. Finally, since both objectivity of ratings and objectivity of ability estimates are well achieved, test users will have more confidence to adopt CR items in their testing situations.

References

Abedi, J., & Baker, E. L. (1995). A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. *Educational and Psychological Measurement, 55*, 701-715.

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard and M. Wilson, (Eds.), *Objective measurement: Theory into practice*. Vol. 3. Norwood, NJ: Ablex.

Adams, R. J., & Wilson, M. R., & Wang, W. (in press). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*.

Adams, R. J., & Wilson, M. R., & Wu, M. L. (in press). Multilevel item response modeling: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*.

Akaike, H. (1977). On entropy maximization principle. In P. R. Krischnaiah (Ed.), *Applications of statistics*. New York: North Holland.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46*, 443-459.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: SAGE.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association, 76*, 341-353.

Engelhard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Engelhard, G. Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*, 56-70.

Goldstein, H. I. (1987). *Multilevel models in educational and social research*. London: Oxford University Press.

Laird, N. M., & Ware, H. (1982). Random-effect models for longitudinal data. *Biometrics, 38*, 963-974.

Lavingueur, S., Tremblay, R. E., & Saucier, J. F. (1993). Can spouse surport be accurately and reliably rated? A generalizability study of families with disruptive boys. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 34*, 689-714.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear models. *Journal of the Royal Statistical Society, B, 34*, 1-41.

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics, 19*, 171-200.

Lunz, M. E., & Stahl, J. A. (1990). *Severity of grading across time periods*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Lunz, M. E., & Stahl, J. A. (1990). The effect of rater severity on person ability measure: A Rasch model analysis. *American Journal of Occupational Therapy, 47*, 311-317.

Lunz, M. E., Stahl, J. A., Wright, B. D., & Linacre, J. M. (1989). *Variation among examiners and protocols oral examinations*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Lunz, M. E., Wright, B. D., Stahl, J. A., & Linacre, J. M. (1989). *Equating practical examinations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement, 54*, 3-7.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention, 18*, 34-47.

Rasch, G. (1960/1980).*Probabilistic Models for Some Intelligent and Attainment Tests*. Copenhagen: Danmarks Paedogogiske Institut.

Rosenberg, R. (1973). Linear regression with randomly dispersed parameters. *Biometrika, 60*, 61-75.

Wang, W. (1994). *Implementation and application of the multidimensional random coefficients logit model*. Unpublished doctoral dissertation of the University of California at Berkeley.

Wang, W., & Wilson, M. R. (1996). Comparing multiple-choice-items and performance-based items using item response modeling. In G. Engelhard and M. Wilson, (Eds.), *Objective measurement: Theory into practice*. Vol. 3, Norwood, NJ: Ablex.

Wang, W., Wilson, M. R., & Adams, R. J. (1995). *Item response modeling for multidimensional between-items and multidimensional within-items*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wang, W., Wilson, M. R., & Adams, R. J. (in press). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice*. Vol. 4, Norwood, NJ: Ablex.

Wilson, M. R. (1992). The partial order model: An extension of the partial credit model. *Applied Psychological Measurement, 16*, 309-325

Wilson, M. R., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*, 51-72.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wu, M., Adams, R., & Wilson. M. (1995). *MATS: Many-aspect test software*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.